CLAIMS:

1. A method of aggregating data comprising the steps of:

receiving data from a plurality of sources;

cleaning the received data, whilst maintaining an audit trail of any changes made to the data in the cleaning step;

creating a data set comprising the cleaned data and the audit trail; and

generating output data using said data set.

2. A method according to claim 1 comprising the further step of standardising the format of the received data before the cleaning step.

3. A method according to claim 1 comprising the further step of splitting the standardised data into respective data types before the cleaning step.

4. A method according to claim 1 in which the audit trail is performed at sub-field level so that there are audit entries in respect of every part of every field that has been modified.

5. A method according to claim 1 in which the audit trail comprises a measure of the quality of the data in said data set.

6. A method according to claim 1 in which the cleaning step is carried out independently in respect of some or all of the respective data types.

7. A method according to claim 6 in which the respective data types comprise names and addresses, and the cleaning step is applied to names and addresses included in the received data.

8. A method according to claim 6 in which the respective data types include at least one of: dates; reference numbers; telephone numbers; e-mail addresses and cleaning is carried out in respect of any one or any combination of these other data types.

9. A method according to claim 1 in which the cleaning step comprises the step of standardising the respective data against a predetermined standard.

10. A method according to claim 9 in which the predetermined standard comprises a predetermined list.

11. A method according to claim 10 which is such as to allow a user to select at least one list against which data is to be standardised.

12. A method according to claim 1 in which the cleaning step comprises standardising the data through the application of rules.

13. A method according to claim 12 which is such as to allow a user to select at least one rule which is applied to the data in the cleaning step.

14. A method according to claim 12 in which the rules are used to at least one of: change the data to a standardised form, correct data, and complete data.

15. A method according to claim 1 in which standardisation against a list is performed in combination with standardisation through rules.

16. A method according to claim 1 in which the cleaning step comprises an automated cleaning process which is intelligent such that it learns from decisions made by human intervention.

17. A method according to claim 1 comprising the further step of matching data records in said data set which relate to a common entity and which originate from respective distinct data sources.

18. A method according to claim 17 in which the step of matching data

records comprises the step of comparing a plurality of data items in respective data records to decide whether the data records relate to a common entity.

19. A method according to claim 18 in which at least one threshold level of similarity between data items is specified, such that the threshold must be met or exceeded before a match is determined.

20. A method according to claim 17 in which decisions on matching are governed by a set of matching rules which specify a plurality of matching criteria at least one of which must be met before a match can be determined.

21. A method according to claim 20 in which each matching criterion identifies at least one predetermined type of data item and at least one similarity threshold.

22. A method according to claim 17 in which the step of matching data records comprises the step of updating the audit trail so as to keep a record of matches made in the matching step.

23. A method according to claim 17 in which an output of the matching process is used to modify the cleaning step.

24. A method according to claim 1 in which the method comprises the further step of de-duplication of data in said data set.

25. A method according to claim 24 in which the step of de-duplication of data comprises the step of updating the audit trail so as to keep a record of changes made to the data set in the de-duplication step.

26. A method according to claim 1 in which the cleaning step is performed iteratively.

27. A method according to claim 17 in which the matching step is performed iteratively.

28. A method according to claim 24 in which the de-duplication step is performed iteratively.

29. Apparatus arranged under the control of software for aggregating data by:

receiving data from a plurality of sources;

cleaning the received data, whilst maintaining an audit trail of any changes made to the data in the cleaning step; and creating a data set comprising the cleaned data and the audit trail.

30. Apparatus according to claim 29 which is further arranged for generating output data using said data set.

31. Apparatus according to claim 29 which is arranged to output a query notification when unable to automatically clean a data item.

32. Apparatus according to claim 31 which is arranged to, allow input of a decision to resolve the query, and complete the cleaning step for that data item based on that decision.

33. Apparatus according to claim 29 which is arranged to learn from a decision input to resolve a query to aid in the cleaning of future data items.

34. A computer program product comprising at least one data carrier carrying a computer program comprising code portions that when loaded and run on a computer cause the computer to carry out a method according to claim 1.

35. A computer program product comprising at least one data carrier

carrying a computer program comprising code portions that when loaded and run on a computer, arrange the computer as apparatus according to claim 29.

36. A method of aggregating data comprising the steps of:
receiving data from a plurality of sources;
creating a virtual data model of the received data; and
using the virtual data model to generate an aggregated data set.

37. A method of generating a virtual data model representing data held by an organisation in a plurality of distinct data sources comprising the steps of:
receiving data from the plurality of data sources;
cleaning the received data, whilst maintaining an audit trail of any changes made to the data in the cleaning step;
creating a data set, as the virtual data model, comprising the cleaned data and the audit trail.

38. A method of aggregating data comprising the steps of:
receiving data from a plurality of sources;
standardising the format of the received data;
splitting the standardised data into respective data types;
cleaning the split and standardised data, whilst maintaining an audit trail of any

changes made to the data in the cleaning step;

creating a data set comprising the cleaned data and the audit trail; and

generating output data using said data set.